COMP5310 Principles of Data Science
Assignment 1

Student name: Joshua Selinger
Student ID: 312090730
Unikey: jsel6715

Problem

The rise of modern streaming services has resulted in unpreceded choice for music consumers (Meneses, 2012). Paradoxically, this has the effect of making it more difficult for consumers to find new music to enjoy (Schwartz, 2004). Music genres are limited in their ability to convey the full set of qualities that motivate consumers engagement with music (Konecni, 1982). As a result, modern day music recommender systems require more sources of information to provide users with music that will suit their preferences.

The Last.fm 2k dataset has been chosen for this project due to its variety of customer data information. It contains two mediums of interaction between the Last.fm's users and music artists as well as the user's social network of friends on the service. Using all three of these user connections to the service provides a rich representation of the user's interests. It also facilitates the possibility of revealing latent factors that relate to the artists. This should ideally translate to better recommendations than any one individual data source.

This project aims to use the combinations of these data tables to develop latent factors and decomposed representations of the user's interactions. These will then be used to produce an artist recommendations system. Evaluation of these recommendations will then be performed on an unseen held out sample of the dataset.

Data

The dataset contains social networking, tagging and music artist listening information from a set of 2,000 users of Last.fm, an online social media and music logging system. The data was collected in 2011 for the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems and made publicly available on its completion (Cantador, Brusilovsky and Kuflik, 2011). For this project the data was accessed directly from the workshop's website (HetRec 2011).

The raw data was a zip file containing six DAT files totalling 12.7 megabytes uncompressed. Five of the files corresponded to the following data tables: 17,632 artists, 92,834 user-listened artist interactions, 11,946 tags, 186,479 artist-tag assignments and 25,434 user friend connections. The sixth file provided a more granular timestamp of the artist-tag assignment. The six DAT files were read into the R programming language using the tidyverse suite of packages (Wickham et al., 2019). All further analysis and pre-processing was also implemented in R.

The main files of interest were the interaction tables, i.e. user-listened artist, artist-tag assignment and user friend connections. The artist and tagged tables served as a natural language look-up tables for the numerically encoded interaction tables. Data size was not a concern. Thus, to ease further data analysis, the look-up and interaction tables were joined on their respective keys. From this exercise it was revealed that 6.8% of the artists in the data set did not have a look-up name pair. Further analysis revealed this corresponded with the lowest percentages of artist in the dataset based on user listen counts. These artists were not removed from the dataset as future learning algorithms would be performed on the numeric ID rather than the human language lookup. Following the table joins samples of each table were visually inspected using the R view window.

As noted by Grčar, Mladenič, Fortuna and Grobelnik (2005) sparsity of recommender system data is a common occurrence due to the vast amount of recommender items relative to any specific user's interactions. Profiling of the Last FM 2k user-listened artist interactions supported this assertion. The distribution of total listens per artist was extremely right skewed as is evident in the 'Level' pane of figure 1. Rather than remove the outlier counts, a natural log transformation was performed instead. As displayed in the 'Log' pane of figure 1, this transformation appears to be approximately normally distributed. It was decided the span and shape of the log transformation made it a more likely candidate input for any future learning algorithms compared to the skewed level data.

A similar exercise was performed on the user-tag assignments table. Of note in observing the count data calculated for this table was the high cardinality of tags and the right skewed distribution. Qualitatively the topics of music genres, moods and specific song qualities were observed with the most commonly used tags associated with broad genres such as 'rock', 'pop', 'alternative' and 'electric'.

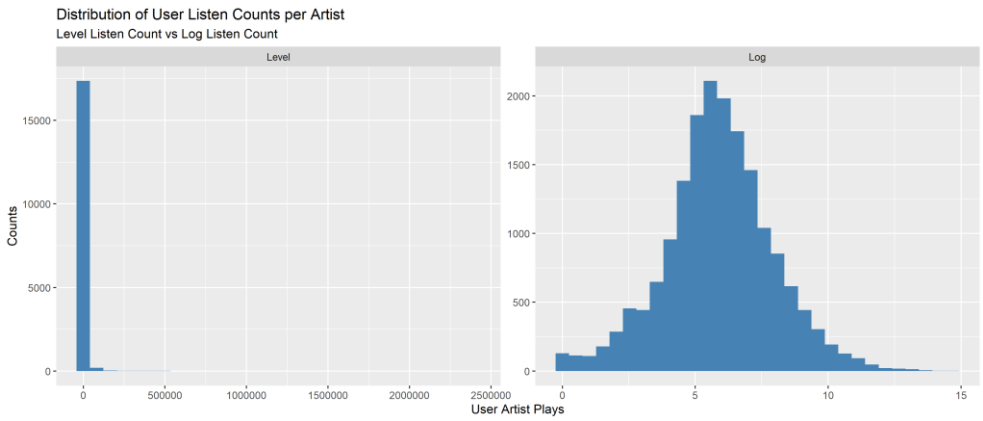

Figure 1 Distribution of User Listen Counts per Artist. Level is highly skewed and sparse, but log transform appears normal.

Of interest in this exploratory stage was what would be the most naïve recommendations as a benchmark. It was hypothesised that this would be to use the most popular artists, whether by listen count or by user count. Ranks with ties were computed for these two cases and compared with a spearman's rank coefficient to assess their similarity. The coefficient evaluated to 0.719, indicating a reasonably strong correlation between the ranks. For easier interpretability and to avoid the noise in the tails of both ranks, a more granular analysis was performed on the top 50 artists by users rank. Appendices figure 2 plots the relationship between the two charts for the top 50 artists by users. Artist with either rank more than two times the other were labelled to differentiate between artists with greater reach of users and those with more listens. Of particularly note was the band Depeche Mode who are ranked second in highest number of listens but ranked 21st in number of users.

A graph theoretic view of the friend relations dataset was taken. Using the R networks package the table, that was made up of every friend partner pair was treated as an edge list. Its dimensions were (25434,2). This list was then computed and plotted as an undirected graph using ggnet2 (Briatte, 2015). Appendices figure 3. demonstrates what appears to be two distinct clusters in the network: one of users with many connected nodes intertwined and another of users with far fewer edges. Further summary statistics and plots were generated of the edge list. Figure 4. demonstrates the distribution of friends per user, which is right skewed with a median of 6.

Proposal

In the next stage, utilising the variety of information in the three interactions inputs, a recommender system is proposed. Its core design is planned to be a matrix factorisation method to reconstruct the user-listened artist interactions. As outlined in Koren, Bell and Volinsky (2009) additional inputs can be incorporated into this reconstruction function. The weights of this function will then be learnt through stochastic gradient descent.

The log transformed user-listened artist interactions are to be decomposed using matrix factorisation collaborative filtering methods, such as those popularised in the Netflix prize (Funk, 2006). This would decompose the log user-listened artist interactions into two matrices ($user\ x\ latent\ factors$) and ($artist\ x\ latent\ factors$). These matrices reconstruct the original user-listened artist interactions via a dot product

The information encoded in the artist-tag assignment and user friend relationship pair will then be added as additional parameters in the reconstruction term. To prevent overfitting a regulariser term is also proposed to be included. This will penalise the magnitude of the included parameters.

Performance evaluation will occur using a train, validation and test split. Hyperparameter is anticipated to occur for the regulariser term of the model. The proposed evaluation metric is RMSE error over the log user-listened artist.

References

Briatte, F. (2015). [online] *https://briatte.github.io* . Available at: https://briatte.github.io/ggnet/ [Accessed on 2 April, 2020]

Cantador, I., Brusilovsky, P., & Kuflik., T. (2011). Last.fm web 2.0 dataset. *2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011), RecSys*.

Funk, S. (2006). [online] *https://sifter.org/* Available at: https://sifter.org/~simon/journal/20061211.html [Accessed on 2 April, 2020]

Grčar, M., Mladenič, D., Fortuna B., & Grobelnik, M. (2005). "Data sparsity issues in the collaborative filtering framework." *In International Workshop on Knowledge Discovery on the Web, pp. 58-76. Springer, Berlin, Heidelberg*.

HetRec 2011 Datatasets. (2011). [online] *http://ir.ii.uam.es/hetrec2011/index.html.* Available at: http://ir.ii.uam.es/hetrec2011/datasets.html [Accessed 30 March. 2020].

Konecni VJ (1982). Social interaction and musical preference. *The psychology of music*, pp 497–516

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30-37.

Meneses, J. P. (2012). About Pandora and other streaming music services: the new active consumer on radio. *Observatorio* (OBS*), 6(1).

Schwartz, B. (2004). The paradox of choice: Why more is less. *New York: Ecco.*

Wickham, H., et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686,

Appendices

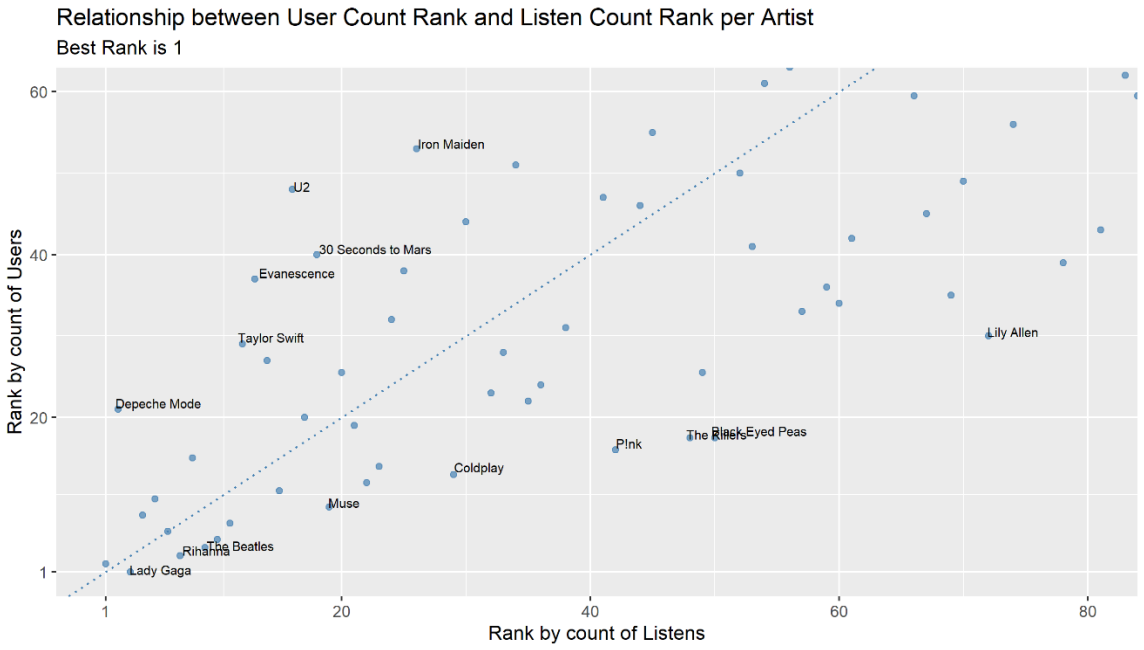

Figure 2: Relationship between User Counts and Play Counts per artist, for top 50 User Artists, 1 is highest rank. The dotted line gradient is one for reference

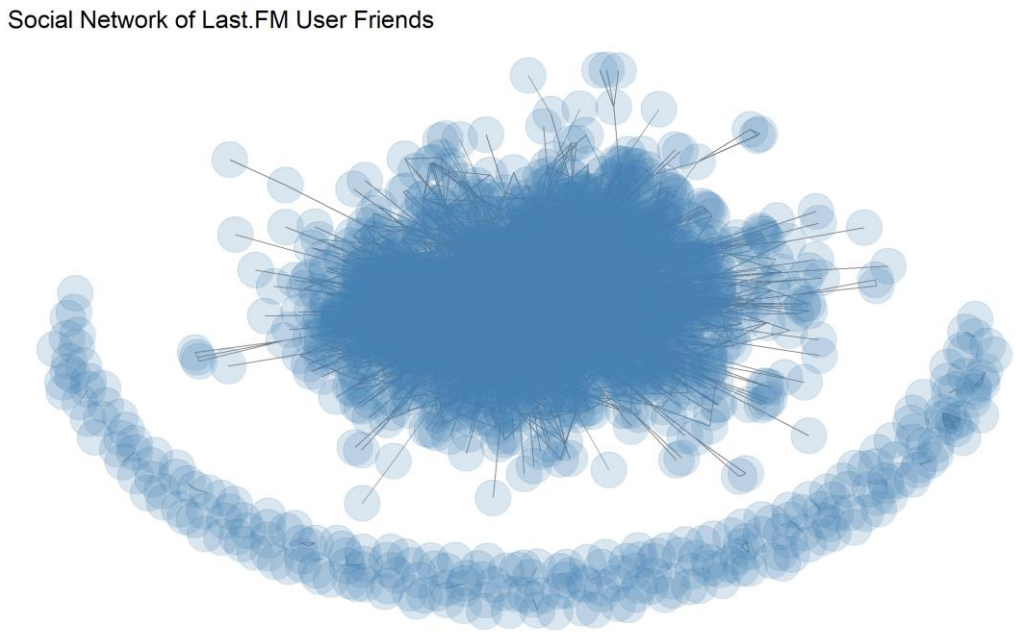

Figure 3: Social Network of Last.FM User Friend Connections

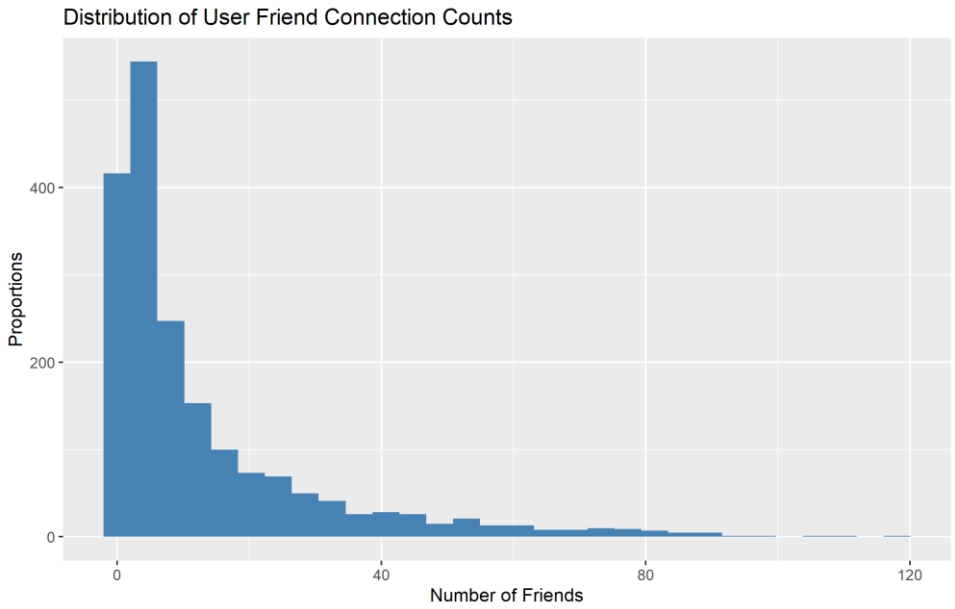

Figure 4: Distribution of User Friend Connection Counts